International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 6, ISSN No. 2455-2143, Pages 103-105
Published Online April - May 2016 in IJEAST (http://www.ijeast.com)

ELASTIC RESOURCES PROVISIONING IN CLOUD COMPUTING SERVICE

Prof. K. S. Balbudhe
Department of IT
PVG's College of Engineering and Technology
Pune, Maharashtra, India

Sukeshani Randive, Pranali Kulkarni, Ujwala Mukane, Shalu Kingrani
Department of IT
PVG's College of Engineering and Technology
Pune, Maharashtra, India

Abstract— **Many internet applications can benefit from an automatic scaling property where their resource usage can be Scale up and Down automatically by the cloud service provider. In this project, we are going to implement the proposed algorithm, but based on cpu utilization as well as memory utilization. We are creating our private cloud on which we are going to deploy the applications and we encapsulate each application instance inside a VM and use virtualization technology to provide instant service. When load on server increases then load will be shifted to another server which is known as load balancing and auto scaling of resource as per requirement. When server reaches to maximum load and cannot handle and more requests, the proposed system is going to provide another healthy instance which handle incoming Request is known as auto scaling. In this paper, we come up with survey of different auto scaling mechanisms as well as resource provisioning techniques.**

Keywords— **Auto scaling, Virtual machine instance, virtualization, class constrained bin packing**

Abbreviations: **VM- Virtual Machine**

I. INTRODUCTION

Cloud computing has grown extremely well in business by effectively providing the world class service to all its users. Cloud computing is a model for enabling on demand network access to a shared pool of configurable computing resource (e.g. network, servers, storage, application and service) that can be rapidly provisioned and released with minimal management effort or service provider interaction. Through cloud computing there is no need to store data on desktop, portable device etc. You can store the data on servers and you access the data on server also you can access the data through internet. The main issue related to cloud computing is load balancing. Load occurs when the number of job increases. Load balancing is a technique in which the workload on the resources is shifted to respective resources on the other node in a network without distributing the running task.

The load to the individual nodes of the collective system to make best response time and also utilization of the resources. We use cloud to balance the load in the public cloud. The parameters are CPU processing speed, queue size, memory size, memory utilization ratio and CPU utilization.

With cloud computing, it is possible to provide cloud services (software/infrastructure platform) as per the requirement. Requirement can fluctuate as per the customer needs and thus the provisioned resources can change. This feature of cloud computing is known as elasticity. This has made cloud popular nowadays. Elasticity is the key concept
in cloud, which allows resource as per request. This elasticity property is gain through scaling up and scaling down the services requested by the customer.

II. EXISTING SYSTEM FOR LOAD BALANCING AND AUTO SCALING MECHANISM

Paper titled "Cloud Auto scaling with deadline and budget Constraints" authored by Min Mao jie Li and jie Li and marty Humphery published in 2012, describes exact constraint related to load balancing and auto scaling which emphasis on the issues such as constraints which can affect the fields people use cloud platform relatively. Auto scaling and load balancing as well .Authors find out matrix when they tried out auto scaling mechanism for VM resources.

1. Requirement of more resource in limited budget. While developing auto scaling mechanism in cloud computing one has to consider user budget while acquiring resources because cloud computing is a platform which offers its users extremely unlimited power of computing task as well as storage mechanism (capacity).

2. Virtual machine instances acquisition time: As auto scaling Provides a way to scale up and scale down the cloud resource any time it does not say that cloud platform scale it very fast in

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 6, ISSN No. 2455-2143, Pages 103-105
Published Online April - May 2016 in IJEAST (http://www.ijeast.com)

(0.0 sec) based On survey, author estimate the VM instance acquisition time to be 10 Minutes. Before the resource is ready for use which is non ignorable otherwise it creates lots of critical condition. Another case is VM shifting down time which takes 2-3 minutes in windows Azure. This case study state that the user has to take care: count in the computing
Power of pending resource; if any instances are in pending status then it means that instances are in ready process. Even though we ignore instances ready status may result in booting of another instances which actually not necessary, this simply implies wastage of money.
Paper titled " Auto Scaling Model for Cloud Computing System" is written by Che-Lun Hung, Yu-Chen-Hu,Kaun-Ching Li published in 2012 describes the novel virtual cluster architecture for dynamic scaling of cloud applications in a virtual cloud computing. An auto scaling algorithm is used for automated provisioning and balancing of virtual machine resources, which is based on active application session and the energy cost is considered in a proposed algorithm. This system has demonstrated that the proposed algorithm is capable to handle sudden load requirements and maintains higher resource utilization with reduced energy cost.
In this paper, auto scaling scenarios are presented to address the automatic scalability of web applications and distributed computing jobs in a virtual cluster on the virtualized cloud computing environment. The cloud computing architecture is constructed with a front end load balancer, a virtual cluster monitor systems and an auto provisioning system. The front-end load balancer route and balance user requests to cloud services which are deployed in virtual cluster. To collect the use of physical resources of each virtual machine in a virtual cluster, the virtual cluster monitor system is used. To dynamically provision the virtual machines based on the number of active session or the use of resources in a virtual cluster, the auto provisioning system is used. The resources are able to release when ideal virtual machines are destroyed. The energy cost can be reduced by removing the ideal virtual machines.

Table- 1: Existing auto scaling and load balancing system

Title of paper	New system introduced	Algorithm used
Cloud Auto scaling with deadline and budget Constraints	Find VM instance Acquisition time	
Auto Scaling Model for Cloud Computing System	Reduce the significant amount of resources	

Paper titled "Resource Provisioning Techniques in Cloud Computing" is written by Bhavani B H and H S Guruprasad published in 2014 describes the resource provisioning techniques which is main challenge in cloud computing. In cloud computing, resources are scale -up and scale-down based on the user demands. Author says that elastic resources provisioning means use of software as per our convenience. As we know that acquiring resources in cloud platform might be costly hence to avoid wastage of money as well time one has to consider resource provisioning techniques. There are two provisioning techniques
1. Static provisioning technique 2. Dynamic provisioning technique many application vendors are uses cloud computing as emerging platform to deploy their applications. Cloud service eliminates the need of setting up infrastructure which takes time. Hence companies most prefer the cloud services. Resource provisioning techniques determines how many amounts of resources are required to execute task which result as saving budget constant. When application vendor deploy their applications on cloud based on need they are classified as:
1. Static provisioning: This is used for applications which is generally unchanging demands.
2. Dynamic provisioning: In case where requirements are not fixed or demands of applications may change or vary on time then this technique is used. Dynamic provisioning has been suggested by virtualization where user can acquire and release resources as per their need. Parameters for resource provisioning are response time, fault tolerance, Revenue maximization.

Table- 2 Comparison of some resource provisioning techniques

Resource provisioning Technique	Metrics	Challenges
Provisioning of request for virtual machine sets with placements constraints in IaaS cloud	Provides effective mean of VM to PM mapping	No practical medium to large problem
Risk aware provisioning and resource aggregation based consolidation of VMs	Reduce the significant amount of resources	Focus on only CPU utilization
Optimal resource provisioning for cloud computing	Efficiently provides resources for SaaS user Only	Applicable only for SaaS Platform

III. CONCLUSION

Considering the growing importance of cloud, finding new ways to improve the cloud services is an area of corner and research focus. We have surveyed various load balancing and auto scaling techniques for cloud computing. The main purpose of load balancing is to satisfy the customer requirements by distributing load dynamically among the nodes and to make maximum resource utilization by the total load to individual. When load is necessary in cloud computing so we have discussed all the existing techniques for load balancing and auto scaling and we have also discussed the virtualization and cloud computing.

IV. REFERENCE

[1] M. Mao; J. Li; M. Humphrey. "Cloud auto-scaling with deadline and budget Constraints," Grid Computing (GRID), 2010 11th IEEE /ACM International Conference , pp.41-48, 25-28 Oct. 2010.

[2] R. Cushing, S. Koulouzis, A. Belloum, M. Bubak, "Predictionbased Autoscaling of Scientific Workflows," 2011 11th IEEE/ACM International Conference , pp. 423–430 Dec. 2011.

[3] T. C. Chieu, A. Mohindra, A. A. Karve, A. Segal. "Dynamic Scaling of Web Applications in a Virtualized Cloud Computing Environment," e-Business Engineering, 2009. ICEBE '09. IEEE International Conference on ,pp.281-286, 21-23 Oct. 2009.

[4] S. Venugopal, L. Han; P. Ray. "Auto-scaling emergency call centres using cloud resources to handle disasters," Quality of Service (IWQoS), 2011 IEEE 19th International pp.1-9, 6-7 June 2011.

[5] J. Zhu, Z. Jiang, Z. Xiao, and X. Li, "Optimizing the performance of virtual machine synchronization for fault tolerance," IEEE Trans. Comput., vol. 60, no. 12, pp. 1718–1729, Dec. 2011.

[6] Lei Shi, Bemard Butler, Dmitri Botvich and Brendan Jennings, "Provisioning of request for virtual machine sets with placements constraints in IaaS cloud", 2013 IFIP/IEEE International Symposium on Integrated network management (IM 2013): Mini conference, pp 499-505, May 2013.